An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM

José Iria **Neil Ireson** **Fabio Ciravegna**
Department of Computer Science
The University of Sheffield
United Kingdom
{j.iria, n.ireson, f.ciravegna}@dcs.shef.ac.uk

Abstract

This paper investigates the incorporation of diverse features in boundary classification algorithms for IE using SVM. Our study reveals that the use of rich data resources greatly contributes to the performance of IE systems and it is more likely to explain the differences in performance reported by several systems than the design decisions relative to the learning model. Evaluation of our system shows an improvement over the state-of-art on a standard dataset using the same data resources but a much simpler learning model than the previously best-reported system.

1 Introduction

Information extraction (IE) is the task of identifying relevant fragments of text in documents. Examples of IE tasks include identifying the speaker featured in a talk announcement or finding the proteins mentioned in a biomedical journal article.

In recent years, several IE algorithms based on standard machine learning algorithms, as opposed to specialized algorithms for IE, have been proposed. Many of these algorithms (Finn and Kushmerick 2004, Li et al. 2005, Gliozzo et al. 2005) are based on support vector machines (SVM) (Cristianini and Shawe-Taylor 2000). The wide adoption of SVM for IE may be justified by the fact that SVM achieves state-of-the-art performance on many other related areas, including named entity recognition (Mayfield et al. 2003), and because there are several readily available fast and robust implementations of the algorithm. This new breed of IE algorithms based on SVM has outperformed the former state-of-the-art in IE (Ciravegna 2001, Freitag and Kushmerick 2000) on several datasets. They all adopt the common "boundary classification" formalization (Freitag and Kushmerick 2000), in which IE becomes the task of classifying every boundary between any two tokens in the document as either the starting position of a fragment to extract, the ending position of a fragment, or neither.

Developing IE solutions based on off-the-shelf ML tools invites to a clear separation between data resources and learning algorithm employed. It allows the IE expert to concentrate on exploring the corpus and linguistic resources to obtain features for the classification task, rather than on the details of the learning machine. Consequently, we claim that the differences between state-of-the-art algorithms for IE reside mostly in the kind of data resources they use and how such resources are transformed into a collection of instances for learning.

The contribution of this paper is threefold. First, we present a study that investigates the incorporation of diverse features in boundary classification algorithms for IE using SVM. The study indicates that the characteristics of the corpus should be taken into account when choosing features for learning. Most importantly, it reveals that the use of rich data resources greatly contributes to the performance of an IE system and it is more likely to explain the differences in performance reported by several systems than the design decisions relative to the learning model. Second, we introduce and describe a software framework designed to allow quick prototyping of IE systems. The framework allows great flexibility in the design and development of IE systems adapted to given problem domains by not requiring, and even discouraging, early commitments on the design decisions. Finally, we present the results of evaluating our system on two standard datasets for IE. Drawing from the conclusions of our experimental study, our system system reports a slight improvement over the previously best-reported system on one of the

datasets, and performs competitively on the other dataset.

The rest of the paper is organized has follows. The following section describes related work. Section 3 introduces the software framework and the IE system used to perform the experiments, which is based on the framework. Section 4 describes in details the set of experiments performed and their motivation. After that, the experimental setup and datasets used in the experiments are described. Section 6 provides an analysis of the results obtained. Section 7 provides a higher-level discussion of the results. Applying the best knowledge gathered from the results of the experimental study, Section 8 presents the comparison with the state-of-the-art. Finally, we outline future work and conclude the paper.

2 Related Work

This paper is concerned with the class of state-of-the-art IE systems that may be characterized as following a boundary classification learning model and using SVM as their learning algorithm. A boundary classification model typically uses a set of binary classifiers to classify boundaries as the start/end of relevant text fragments – usually referred to as slot fillers. Positive instances for the start/end classifier for a given slot become negative instances for all other classifiers.

Finn and Kushmerick (2004) introduced a variant to the usual boundary classification approach which makes use of a two-level ensemble of classifiers. The approach takes advantage of the fact that high-confidence predictions for the start of a fragment are an indication of its end in the nearby text, and vice-versa. In the first level, their approach uses high-precision classifiers so as to spot individual start or end of fragments. On the second level, their approach uses high-recall classifiers, but restricted to the vicinity of the individual start/end already predicted by the first level classifiers. Their SVM-based IE system implementing the multi-level approach to boundary classification, called ELIE, showed state-of-the-art performance on standard datasets.

Li et al. (2005) describe in great detail their SVM-based system for IE, called GATE-SVM. One distinctive feature of the system is that it uses a variant of SVM, the SVM with uneven margins, which the authors show to be particularly helpful for small datasets. Another interesting feature of GATE-SVM is that it uses a weighing scheme for the token features according to distance of the token to the boundary in the text. Their system obtained very good results in the international competition "Evaluating Machine Learning for Information Extraction" organized by the PASCAL network (Ireson et al. 2005).

Gliozzo et al. (2005) uses a SVM-based approach and system (called SIE) similar to Li et al. They propose instance filtering as a preprocessing step for classification-based IE systems. The goal of instance filtering is to reduce the skew of the class distribution and the dataset size by eliminating negative instance while preserving positive instances as much as possible. They reported an impressive reduction in computation time required to train and classify, while maintaining (and sometimes improving) prediction accuracy.

3 The T-Rex Framework

The T-Rex framework[1] (Iria 2005) was designed and developed with the main goal of allowing quick prototyping of IE systems adapted to a given domain. The framework aims to be general enough to support a variety of entity extraction and relation extraction algorithms.

Amongst the distinctive characteristics of T-Rex are a clear separation between data representation and learning algorithm, complete parametrization available to user in a declarative way, including user customizable model for data representation and user customizable model for classification, support for classification of multiple types of objects, support for classification using multiple views, support fine-tuning for specific classes, and explicit mechanisms to adjust the memory/speed trade-off. In this paper, we focus in the aspect of clearly separating data representation and learning algorithm, as it is more relevant to our claim.

3.1 Data Representation

Even systems relying on identical learning algorithms and the general boundary classification model face a plethora of design decisions such as how to obtain features from the tokens in the vicinity of the boundary, which features derived from external resources and processors (e.g. parts-of-speech, ontologies) to use, whether some features are more relevant than others, which features should be combined and so on.

While in many IE systems data representation and algorithm are tightly coupled, the T-Rex framework provides mechanisms to integrate corpus and external resources of different levels

[1] Available for download at http://tyne.shef.ac.uk/t-rex/

of complexity, ranging from simple resources like gazetteer lists to layout information from HTML to domain ontologies, into a uniform graph-based representation. The advantage of an uniform representation is that access and querying from the point of view of the leaning algorithm is standardized. This allows for declarative methods for accessing the representation. Concretely, access to the representation is organized around three concepts: graph walks, cost models for relation traversal and feature sensors.

A graph walk is a function operating over a set of nodes in the graph by posing conditions on the relation traversal. Graph walks are defined by a grammar that includes composition operators like set intersection, set union, node set replacement and walk repetition, which are built on primitive relation traversal operators. The cost model specifies the cost of traversing relations. The result of applying a graph walk to the graphical data representation is a set of subgraphs matching the conditions on the relations. Feature sensors are employed by learning algorithms to obtain features from the data representation. Sensors extensively use graph walks to access the graph structure and collect subgraphs that can be transformed into features in a number ways, according to the particular sensor used.

4 A Study on Boundary Classification Algorithms

The experimental study here presented investigates several variants of the general boundary classification model. The study intends to clarify the contribution of different features to the overall performance of SVM-based boundary classification systems. The variants are obtained in a number of ways as described in the following subsections.

4.1 Effect of Combining Feature-sets

Typical external data resources and processors used in IE include sentence splitters, tokenizers, parts-of-speech taggers and gazetteers. This experiment will show the effect of combining these resources and their contribution to the different slots to extract. The experiment will combine four kinds of token-related features: the token string, the token part-of-speech, the token orthography (or word shape), and categories for the token looked up in a gazetteer. These feature-sets are denoted by S, P, O, and G respectively. The data resources used for this experiment are the default ones provided by the NLP tools chosen (see Section 5).

4.2 Effect of Quality of Features-sets

This experiment takes the resources of the previous experiment and tries to improves them. The parts-of-speech are organized in a tree structure where a part-of-speech tag can have a parent tag, e.g., VBD, VBN and VBZ tags may have a more generic VB parent tag. When a tag is inserted as a feature, its ancestors up to the root of the tree are also inserted. This potentially helps the learning machine to generalize better. The orthographic categories for this experiment were augmented with specific categories for one and two letter words, words containing special characters and acronyms, inspired by LP^2 (Ciravegna 2001). Moreover, the orthography is also organized hierarchically in a similar manner to the parts-of-speech. The gazetteer used in this experiment is the gazetteer used in Finn and Kushmerick (2004). This gazetteer includes roughly the same categories as the one used in 4.1, but contains many more entries, particularly related to person's first and last name. Concretely, it contains several tenths of thousand entries for first and last name, whereas the gazetteer in 4.1 only contained a few hundred entries for first name. In contrast, the gazetteer used in this experiment contains fewer categories for date and time than the one in 4.1.

The feature-sets for this experiment are denoted by S', P', O', and G' respectively

4.3 Effect of Space and Newline Tokens

The effect of space and newline tokens is dependent on the nature of the dataset and of the slot to extract. This experiment explores three variants in the way the corpus is preprocessed: removing all space and newline tokens; removing just space tokens, keeping newline tokens; and keeping all token types.

4.4 Effect of Token Window Length

Boundary classification typically involves taking tokens in the vicinity of the boundary to generate features, forming a so-called “window” of tokens around the boundary. This experiment analyzes the impact of the chosen length for the window in the performance of the system.

4.5 Effect of Feature Selection

Gliozzo et al. (2005) have shown that instance selection is technique able to greatly reduce the complexity of the learning problem while maintaining accuracy. Inspired by their work, this experiment takes standard feature selection metrics widely used in text categorization and applies them in the context of the boundary classifica-

tion in IE. In contract with the text categorization field where feature selection has been widely studied, little is known about the effects of using feature selection in IE.

The feature selection metrics used in this experiment are cross-entropy, information gain, frequency and a random baseline metric. For details on the metrics refer to (Sebastiani 1999).

5 Experiments

5.1 Datasets

The experiments were performed using two standard benchmark datasets: the seminar announcements ("SA") corpus (Freitag 1998) and the workshop call for papers ("WCFP") corpus (Ireson et al. 2005). SA consists of 485 seminar announcements from Carnegie Mellon University detailing upcoming seminars. Each seminar is annotated with slots speaker, location, start-time and end-time. WCFP is a corpus recently created for the international PASCAL Challenge entitled "Evaluating Machine Learning for Information Extraction". It consists of 1100 workshop call for papers, 600 of which were annotated. From those, 200 were never released by the organizers of the challenge. These experiments use the publicly released 400 annotated documents. The corpus includes 11 slots such as workshop and associated conference names, acronyms, locations and dates, and the deadlines for paper submission, notification of acceptance and camera-ready version. The experiments use a random 50:50 split of the SA dataset and a random 75:25 split of the WCFP dataset.

5.2 Scoring

The results are reported using the F1-measure, which is the harmonic mean of precision and recall, i.e., F1 = (2 x p x r)/(p + r), where *p* (precision) is the percentage of correct entities found by the system and *r* (recall) is the percentage of entities in the test set found by the system. A predicted annotation is only considered to be a match if it strictly matches the human-annotated tag, both in terms of its type and its start and end offsets in the document. Concerning averaging of the scores, macro-averaged was used mainly because it was not possible to get micro-averaged results for some of the systems being compared.

5.3 Baseline System

The experiments use a baseline system and modify it according to the variant being tested. The baseline system pre-processes the corpus using the default ANNIE components of the GATE[2] system, namely the default tokenizer, parts-of-speech tagger and gazetteer. In the baseline system no tokens are discarded, not even space tokens. The feature-sets are encoded as features in the same way as described in Finn and Kushmerick (2004) and Li et. al (2005). The default window length is 10 tokens to each side of the boundary. The SVM implementation used in all experiments is SVMLight[3] (with parameters j=2, c=0.075 for SA and j=10, c=0.05 for WCFP, optimized by cross-validation). Feature selection is always performed on each binary classifier's instances separately. At the end of the classification process, the predictions for start and end of the text fragments are paired by (1) recursively enumerating all possible pairs for each given segment of the document (2) calculating a score for each possible subset of the superset of pairs, based on the sum of classifier confidence measures for the individual predictions and (3) selecting the set of pairs that maximizes the confidence score.

6 Analysis of Results

The results obtained for the experiment described in subsection 4.1 are shown in Tables 1 and 2. Note that only a few of the slots for WCFP are shown, due to space limitations. In all experiments, the slots chosen for WCFP are those which exhibit higher sensitivity to changing features.

Feat.	***location***	***etime***	***stime***	***speaker***	***macro***
O	0	12.3	52.1	0	16.1
G	0	76.5	71.6	7	38.78
P	37	84.7	82.6	42.78	61.77
S	82.3	95.9	**94.8**	53.7	81.7
POG	72.8	95.4	91.9	70.54	82.7
SG	83.2	94.7	94.5	72.6	86.2
SO	85.9	**96.4**	94.3	69.1	86.4
SP	86.2	96.2	94.2	70.9	86.9
SPO	**86.7**	96.2	94.2	72	87.3
SOG	85.7	94.9	94.7	77.8	88.3
SPG	86	95.9	94.5	78.5	88.7
SPOG	86.2	95.9	94.5	**78.8**	**88.9**

Table 1. The effect of different feature-sets on the SA dataset. The best F-measures for each slot are highlighted.

[2] http://www.gate.ac.uk
[3] http://svmlight.joachims.org/

Feat.	*wacr*	*wloc*	*wdat*	*whom*	*macro*
O	29.6	0	0	4.1	6.9
G	0	42.1	59.2	1.3	19.2
P	53.2	38.1	53.3	49.4	41.3
POG	57.3	53.3	66.5	54.7	50.3
SP	69.3	60.3	67.4	58.5	60.4
S	69.6	55.3	69.7	60.5	60.6
SG	67.7	63.6	70.4	60.4	60.8
SOG	70.5	62.6	71.1	60	60.9
SPO	**71.7**	58.1	68.5	58.1	61
SPOG	69.6	63.2	**71.5**	58.2	61.1
SO	71.6	58.9	70.3	**61.3**	61.4
SPG	71.6	**63.9**	70.6	58.4	**62.1**

Table 2. The effect of different feature-sets on the WCFP dataset. Only slots workshop acronym, location, date and homepage are shown.

Table 1 shows the contribution of each feature-set to the results obtained for SA. Roughly, the more feature-sets used the better the results, confirming the rule of thumb "the more features the better" well-known due to robustness of SVM to noisy data.

The inclusion of the gazetteer feature-set boosts the results of the *speaker* slot. For *location*, the addition of gazetteer seems to always have a slightly negative effect. Slots *stime* and *etime* achieve very good results with no more than the plain token string features.

Table 2 shows results for WCFP. The rule of thumb "the more features the better" still seems to apply, although not as clearly as in the experiments for SA. For WCFP the additional feature-sets on top of token string seem to carry less information than in the case of SA, since using token string alone achieves impressively good results (compare with the best feature-set combination). The slots workshop *location* and *date* score highest whenever the gazetteer is present. Workshop *homepage* seems to obtain slightly worse results whenever parts-of-speech or gazetteer are added.

The improvements on the *speaker* and workshop *location* and *date* given by the use of a gazetteer are as expected. The exception is *location* of SA, which may be explained by the fact that seminar locations consist of room names/numbers rather than city/country names.

Feat.	*location*	*etime*	*stime*	*speaker*	*macro*
O'-O	**54.2**	**76.9**	**33.8**	**41.3**	**51.5**
P'-P	8.4	3.3	2.1	3.9	4.4
G'-G	0	**-16.2**	**-25.8**	**61.8**	5
SP'-SP	-0.4	0.3	-0.7	-0.3	-0.3
SO'-SO	1.1	-1.5	-1	0.9	-0.2
SP'O'G -SPOG	-1.3	-1.8	-0.8	-2.5	**-1.6**
SPOG'-SPOG	0.8	1	0.2	**6.82**	**2.2**

Table 3. The effect of the quality of the feature-sets for SA . Selected results are highlighted.

Feat.	*wacr*	*wloc*	*wdat*	*whom*	*macro*
O'-O	**22.1**	**40**	**54.3**	**52.6**	**33.4**
P'-P	-1	-1.6	-0.9	1	-0.2
G'-G	0	-3	**-39.9**	-1.3	**-7.4**
SP'-SP	0.5	-0.3	0.9	0.3	0.4
SO'-SO	1	2.5	0	-1	0.5
SP'O'G -SPOG	0.7	2.2	-0.9	-1.4	0.5
SPOG'-SPOG	1.1	-1	2.9	-3.9	0.2

Table 4. The effect of the quality of external resources on the WCFP corpus. Some selected results are highlighted.

The results for the experiment described in subsection 4.2 are shown in Tables 3 and 4. Clearly, O' and P' perform much better in isolation than their counterparts O and P. However, when used together with other feature-sets their informative value is actually not that great. Overall, the use of O' and P' does not constitute a clear improvement, as the results differ on the two datasets. Also, note that gazetteers G and G' perform very differently. On slots related to date and time, e.g. *etime*, *stime*, *wdat*, G performs better, while on slots related to people's names, namely *speaker*, *G'* performs better. This is easily explained by considering the characteristics of the gazetteers (see subsection 4.2).

The results for the experiment described in 4.3 are shown in Tables 5 and 6. In both datasets a clear improvement was obtained by removing space tokens only. For these datasets (in fact most datasets) the semantics of spaces has little predictive value, and treating spaces as separate tokens can adversely influence the ability of the learning machine to generalize. Most importantly, the presence of newline tokens seems to have great influence on the results for both datasets.

SA	*location*	*etime*	*stime*	*speaker*	*macro*
A	82.9	96.1	93	72.3	86.1
B	85.8	**97.4**	94.1	**79.4**	**89.2**
C	**86.3**	95.9	**94.6**	78.8	88.9

Table 5. The effect of space and newline tokens on the SA dataset. A=removed spaces and newlines, B=removed spaces only, C=nothing removed.

WCFP	*wloc*	*wdat*	*whom*	*wnam*	*macro*
A	63.9	72.2	**60.7**	55.7	62.6
B	**71.8**	**74.4**	57.6	**65.9**	**65**
C	63.2	71.5	58.2	59.9	61.1

Table 6. The effect of space and newline tokens in the WCFP dataset. A=removed spaces and newlines, B=removed spaces only, C=nothing removed.

The results also hint about the underlying nature of the corpora. In the case of SA, newline tokens seem to be important which is not surprising given the high regularity of the document formatting. In the case of WCFP, slots are much less regular and the negative effect of space tokens in the generalisation capabilities of the learning machine is more noticeable.

The results for the experiment described in 4.4 are shown in Figures 1 and 2. Again only a few selected slots for WCFP are shown in the graph, for the sake of clarity.

There seems to be an optimal window length for each slot in each dataset. Evidently, the token window should be large enough in order to capture useful patterns. But it is somewhat surprising to learn that for windows that are "too large" there is a constant small drop in the accuracy. This can be seen as a sign of over-fitting. There are even some slots, for instance *workshophomepage* and *conferenceacronym*, that reveal a significant drop in accuracy as the window length gets larger and larger. In contrast with comments on the previous experiments, this contradicts the rule of thumb "the more features the better". Roughly, the optimal average window length according to the macro-average of all slots seems to be around 9 for both datasets.

The results for the experiment described in 4.5 are shown in Figures 3 and 4. The behavior of various feature selection metrics is similar in both datasets.

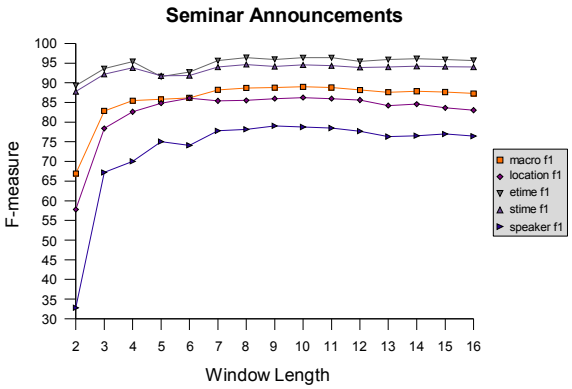

Figure 1. The effect of token window length on the SA corpus. The x axis show the length of the window in number of tokens to either side of the boundary, while the y axis shows the F-measure obtained.

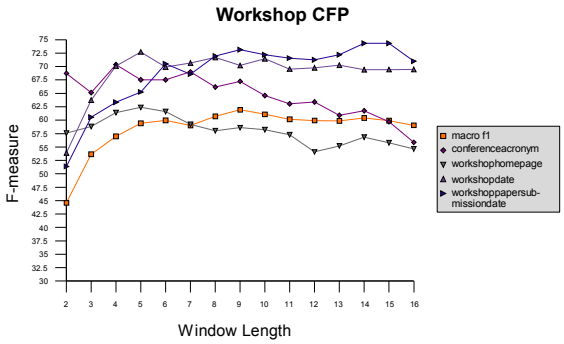

Figure 2. The effect of token window length on the WCFP corpus. The x axis show the length of the window in number of tokens to either side of the boundary, while the y axis shows the F-measure obtained.

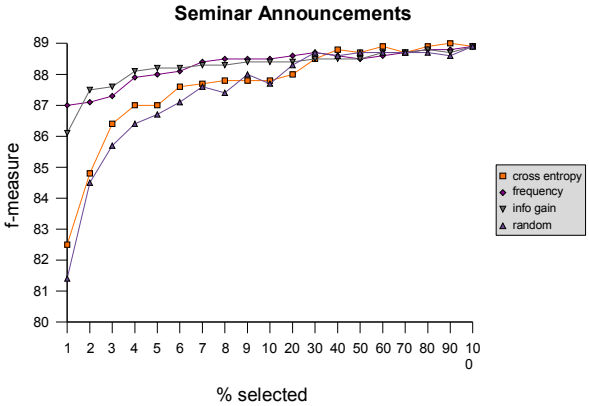

Figure 3. The effect of feature selection on the SA dataset. The x axis show the percentage of features selected, while the y axis shows the F-measure obtained.

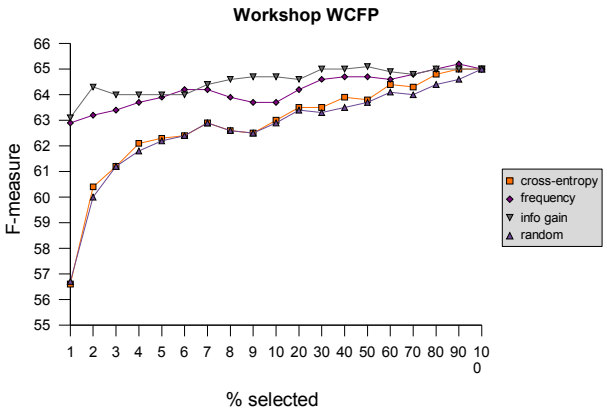

Figure 4. The effect of feature selection on the WCFP datasets. The x axis show the percentage of features selected, while the y axis shows the F-measure obtained.

In general, cross-entropy does not improve much over the baseline random. Information gain and frequency metrics are able to effectively reduce the number of features used to 5 to 10% of the total features with little loss in overall accuracy. For some ranges of relevant features there was a slight improvement over the accuracy obtained by the algorithm that does not use feature selection.

7 Discussion

The experiments performed in this paper on two different standard datasets for IE suggest that, even though slightly different combinations of feature-sets provide the best results on each of the datasets, the rule of thumb "the more features the better" may be applied, albeit with some care, in most situations.

The underlying nature of the corpora and the external resources, especially the gazetteer, was evidenced throughout the results. The high regularities in the document formatting of the SA corpus determine that newline tokens should be used. The difficulty in generalizing patterns over the WCFP corpus means that spaces should be avoided. The various slots related to dates and times in WCFP benefit from a gazetteer that discriminates well in those categories, while a slot like *speaker* boosts the results on SA when a gazetteer that contains substantial data about person's names.

The results obtained for feature selection show that even the simple frequency metric can greatly reduce the number of features with no significant loss in terms of accuracy. However, in contrast with the application of feature selection to the text categorization problem, where the use of some of the metrics is known to consistently improve accuracy, there was no observed improvement in the accuracy with any of the metrics when applied to the information extraction problem.

8 Comparison with the state-of-the-art

Drawing from the lessons learned from our experimental study, we designed an IE algorithm for comparison with the state-of-the-art. For the SA dataset, the baseline system used in the experiments was modified to remove space tokens only in preprocessing and use the gazetteer used by Finn and Kushmerick (2004). For the WCFP dataset, the baseline system used in the experiments was modified to remove space tokens only in preprocessing. No adjustment was required on the token window length nor was any feature selection method adopted.

Care was taken to ensure the experiments were reproduced exactly as the original authors described them - see concerns about the comparability of experiments in IE in Lavelli et al. (2004). Therefore, for the SA dataset, we used the same random 50:50 splits repeated ten times and and the exactly the same gazetteer as used by Finn and Kushmerick (2004) in their experiments. For the WCFP dataset, we used the same standard feature-sets and the same 4-fold cross-validation splits imposed by Ireson et al (2005) for evaluation of the system that participated in the international competition.

Table 7 compares our system with the state-of-the-art for the SA dataset, while Table 8 compares our system with the state-of-the-art for the WCFP dataset.

On the SA dataset, our system reports a small improvement over the previously best-reported results. Note that *speaker* is usually considered the most difficult slot to extract for this dataset. The inferior results obtained by the Gate-SVM system may be explained by the fact that it uses a data-poorer gazetteer, according to the discussion in the previous sections.

SA	***Ours***	***ELIE***	***GATE-SVM***
location	84.9	**85.9**	81.3
stime	93.1	90.2	**94.8**
etime	93.6	**94.6**	92.7
speaker	**85.9**	84.9	69
macro-avg	**89.4**	88.9	84.5

Table 7. Comparing our system with the state-of-the-art on the SA dataset. Macro-averaged F-measures of all slots are presented.

Our system compares reasonably well against the other state-of-the-art SVM-based systems on the WCFP corpus, obtaining better results than those obtained by ELIE. Most importantly, note that our system is considerably simpler than ELIE in the sense that it does not require a multi-level classification approach in order to achieve better results on the two corpora.

The authors of the GATE-SVM system mention feature weighing according to distance to the boundary as a technique that somewhat improves the results on this dataset. We have tried to replicate those improvements without succeeding so far.

WCFP	***Ours***	***ELIE***	***GATE-SVM***
wname	58.1	55.5	**60.6**
wacronym	66.6	68.3	**69.7**
wdate	**78.2**	70.9	76.8
whomepage	63.8	62.8	**68.5**
wlocation	**67**	55.5	66.9
wpsubdate	77.4	70.5	**79.3**
wnotifdate	78.9	71.9	**80.9**
wcamedate	72.8	68.7	**75.9**
cname	62.3	**66.5**	66
cacronym	60.6	**69.1**	66.1
chomepage	29.7	**43.3**	33.1
macro-avg	65	63.9	**67.6**

Table 8. Comparing our system with the state-of-the-art on the WCFP dataset. Macro-averaged F-measures of all slots are presented.

9 Conclusion

In this paper, we presented experimental results on several boundary classification algorithms for IE using SVM. The experiments allowed us to derive some conclusions about the specific characteristics of the corpora. We used that knowledge as input into the design of an IE system competitive with the state-of-the-art.

The comparison with the state-of-the-art confirmed that the use of rich data resources greatly contributes to the performance (or lack thereof) of the systems and, in this particular case, explains the differences in performance reported by the various systems in the literature on one of the benchmark datasets.

In future work, we plan to perform other studies regarding the several other aspects of the boundary classification model for IE, including evaluating the incorporation of other external data resources into the IE process and alternative ways to derive features from those resources. We also plan to concentrate on the tag pairing problem and perform comparative studies on several ways to perform tag pairing.

References

Ciravegna, F. 2001. *Adaptive information extraction from text by rule induction and generalisation*. In Proceedings 17th Int. Joint Conference Artificial Intelligence.

Cristianini, N. and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines (and other kernel based methods)*. Cambridge University Press.

Finn, A. & Kushmerick, N. 2004. *Multi-level Boundary Classification for Information Extraction*. In Proceedings of the 15th European Conference on Machine Learning, Pisa, Italy.

Freitag, D. 1998. *Machine Learning for Information Extraction in Informal Domains*. PhD thesis, Carnegie Mellon University.

Freitag, D. and Kushmerick, N. 2000. *Boosted wrapper induction*. In Proceedings 17th Nat. Conference Artificial Intelligence.

Gliozzo, A. et al. 2005. *Instance Filtering for Entity Recognition.* In ACM SIGKDD Explorations, special Issue on Natural Language Processing and Text Mining, 7(1), pp. 11-18.

Lavelli, A. et al. 2004. *A Critical Survey of the Methodology for IE Evaluation*. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), pages 1655-1658, Lisbon, Portugal.

Li, Y. et. al. 2005. *SVM Based Learning System For Information Extraction*. In Proceedings of Sheffield Machine Learning Workshop. Lecture Notes in Computer Science. Springer Verlag.

Ireson, N. et al. 2005. *Evaluating Machine Learning for Information Extraction*. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany.

Iria, J. 2005. *Relation Extraction for Mining the Semantic Web*. In Proceedings of the Dagstuhl Seminar on Machine Learning for the Semantic Web, Dagstuhl, Germany.

Mayfield, J. et al. 2003. *Named entity recognition using hundreds of thousands of features*. In Proceedings of CoNLL-2003, pp. 184-187. Edmonton, Canada.

Sebastiani, F. 1999. *Machine Learning in Automated Text Categorization*. ACM Computing Surveys 34.